Naive Bayes: Shakespeare or Malory

Nathaniel Beckemeyer

1 Datasets

I selected two documents for text analysis: Sir Thomas Malory's *Le Morte d'Arthur* and the complete works of William Shakespeare. Both of these documents were made available by Project Gutenberg, and were submitted with this report. I chose these documents because they are moderately sized, and written about 100 (depending on when Shakespeare wrote his various works) years apart, which perhaps facilitates their differentiation. I also thought that it would be fun to ask the clickbait-esque question "Which old-timey writer are you?"

The Malory dataset had 362616 words, and the Shakespeare set had 901325. Words are defined as the items created by applying Python 3's in-built 'split()' function to each document (so they can include punctuation).

2 Experiments

Each document was put through 1000 trials as a list of words. Each iteration, the document was randomly split into a test and train set. The test set consisted of 10 words for each document. I chose a much smaller number of words for the test sets because, for large samples of words, the classification accuracy repeatedly achieved 100%. This behavior is expected because the probability of receiving certain specific terms (such as "Merlin" or "Romeo") increases greatly and the word usage approaches the author's actual word usage.

3 Results

Results are presented in Table 1. Unsurprisingly, the Malory dataset had a higher average accuracy. Because the probability of observing a word was considered to be the proportion of words from a single document that was that word, having fewer words would give the Malory dataset an advantage. Overall, both did very well, having an accuracy above 0.9 for each of them.

Additionally, Table 2 presents some more fun results. Note that the word "Titanic" does not appear in either document, so this is the probability distribution for an unseed word. In fact, these probabilities correspond perfectly to the proportion of the numbers of total words.

Tab. 1: The classification accuracy means and standard deviations for test sets of 10 words from Sir Thomas Malory's *Le Morte d'Arthur* and the complete works of William Shakespeare, averaged over 1000 trials.

Author	μ	σ
Shakespeare	0.929	0.257
Malory	0.993	0.083

Tab. 2: Some fun classification results, based on the entire set of words in each document.

Phrase	p(Shakespeare)	p(Malory)
wherefore art thou	0.592	0.408
heart	0.702	0.298
Merlin	0.006	0.994
Romeo	0.951	0.049
Capulet	0.668	0.332
knight	0.009	0.991
the	0.433	0.567
skull	0.707	0.293
Titanic	0.287	0.713
fairy	0.945	0.055
magic	0.801	0.199
sword	0.170	0.830
Gutenberg	0.670	0.330